心理科学进展 2020, Vol. 28, No. 5, 752–765
Advances in Psychological Science DOI: 10.3724/SP.J.1042.2020.00752

· 研究前沿(Regular Articles) ·

人声加工的神经机制*

伍 可[1,2] 陈 杰[1,2] 李雯婕[1,2] 陈洁佳[1,2] 刘 雷[3] 刘翠红[1,2]

([1]湖南师范大学教育科学学院; [2]湖南师范大学认知与人类行为湖南省重点实验室, 长沙 410081)
([3]北京大学心理与认知科学学院, 北京 100080)

摘 要 人声是人类听觉环境中最熟知和重要的声音, 传递着大量社会相关信息。与视觉人脸加工类似, 大脑对人声也有着特异性加工。研究者使用电生理、脑成像等手段找到了对人声有特异性反应的脑区, 即颞叶人声加工区(TVA), 并发现非人类动物也有类似的特异性加工区域。人声加工主要涉及言语、情绪和身份信息的加工, 分别对应于三条既相互独立又相互作用的神经通路。研究者提出了双通路模型、多阶段模型和整合模型分别对人声的言语、情绪和身份加工进行解释。未来研究需要进一步讨论人声加工的特异性能否由特定声学特征的选择性加工来解释, 并深入探究特殊人群(如自闭症和精神分裂症患者)的人声加工的神经机制。

关键词 人声加工; 特异性; 颞叶人声加工区; 言语加工; 情绪韵律; 人声身份识别
分类号 B842

1 引言

人声和人脸是人类听觉环境和视觉环境中非常重要的刺激, 两者传递了相似的社会相关信息。比如, 它们都传递着言语(speech)信息(由人声的音素和人脸的视位传递)和副语言(paralinguistic)信息(如说话者的性别、年龄、情绪状态) (Belin, Bestelmeyer, Latinus, & Watson, 2011)。越来越多的研究发现人声加工与人脸加工的神经机制存在许多的相似之处, 所以有研究者把人声又称作"听觉人脸" (Belin, 2017; Jiang, Chevillet, Rauschecker, & Riesenhuber, 2018)。人脸加工的神经机制得到了大量深入的研究, 且大量电生理和脑成像研究的证据表明大脑有特定的模块来加工人脸, 即对人脸的加工具有特异性(Besson et al., 2017; Caharel et al., 2011; Kanwisher, McDermott, & Chun, 1997; Navajas, Nitka, & Quiroga, 2017)。近年来, 研究者们采用功能磁共振成像(functional magnetic resonance imaging, fMRI)、近红外信息分析(near-infrared spectroscopy, NIRS)、事件相关电位 (event-related potential, ERP)、脑磁图(magnetoencephalography, MEG)、单细胞记录(single cell recordings)等技术考察了人脑是否也存在特异性的人声加工模块(Agus, Paquette, Suied, Pressnitzer, & Belin, 2017; Belin, Bodin, & Aglieri, 2018; Belin, Zatorre, Lafaille, Ahad, & Pike, 2000; Capilla, Belin, & Gross, 2013; Perrodin, Kayser, Abel, Logothetis, & Petkov, 2015)。

此外, Bruce 和 Young (1986)建立的人脸感知模型提出大脑首先会对人脸的结构特征进行编码, 再对人脸的情绪信息、言语信息、身份信息分别进行分析。与人脸加工类似, 研究者发现大脑在识别出人声之后, 会有三条神经通路分别对人声中的言语、情绪、身份信息进行更加精细的分析(Belin, Fecteau, & Bédard, 2004)。关于人声加工, 大量研究主要探索了人声的言语信息加工, 而忽略了对副语言信息加工的探索(如人声情绪和身份信息)。进化心理学研究表明语言是在人类进化和社会文化发展过程中由非语言发声(如笑声、哭泣声)逐渐演变而来的, 人类对副语言加工的神经机制要早于言语加工的存在(Fischer, 2017; Perrodin et al., 2015; Schroeder, Kardas, & Epley, 2017)。

收稿日期: 2019-07-11
* 湖南省哲学社会科学基金(15YBA263), 湖南省教育厅科学研究项目(18A036)。
通信作者: 陈杰, E-mail: xlxchen@163.com

因此，本文将重点介绍近年来人声加工的认知神经科学研究进展。首先，介绍人声加工的特异性脑机制；然后从人声的言语、情绪和身份信息加工三个方面来阐述人声加工的三条神经通路及理论模型；最后就人声加工的特异性、特殊人群的人声加工及自我声音加工等方面，对未来研究进行展望。

2　人声加工的特异性研究

2.1　脑成像研究的证据

Belin 等(2000)利用 fMRI 技术首次发现对人声有选择性反应的脑区位于颞上沟(superior temporal sulcus, STS)。在实验中，他们让被试聆听人类声音(言语声音，如单词；非言语声音，如笑声、叹息和咳嗽声)和非人类声音(如自然声音、动物声音和机械声音)，结果表明无论人声中是否包含言语，人声都比非人声激活更多的区域，且双侧颞上沟上岸(upper bank of STS)对人声的神经反应最强。随后，研究者将人声和加扰人声(保留了人声的频谱包络但听起来不像人声)以及与人声的频率分布保持一致的噪音进行了比较，发现人声刺激比其他声音引起 STS 的中央区域更大的激活(Belin et al., 2000)。鉴于动物发声与人类发声在声学特征上的相似性，研究者又专门考察了动物发声和人声的脑激活模式，结果仍表明人声相比动物发声对 STS 的前部有更强的激活(Fecteau, Armony, Joanette, & Belin, 2004)。尽管以上研究都证明了 STS 中存在人声的选择性区域，但是这种人声的选择性反应也可能是由于 STS 对人声中一些特定低级声学特征进行了选择性反应。研究者们发现人声比其他声音有更多的谐波结构或更复杂的频谱(Leaver & Rauschecker, 2010; Lewis et al., 2009)。因此，有研究者将乐器声和人声在低级声学特征(如音高、音强、谐波噪声比和频谱轮廓)上进行匹配后，再比较了乐器声和人声的脑激活特征，结果进一步发现颞上回或沟(superior temporal gyru/sulcus, STG/S)对人声有更强烈的选择性反应(Agus et al., 2017)。

脑成像研究表明，类似于视觉皮层的梭状回人脸区(fuisform face areas, FFA)，人类听觉皮层中也存在对人声进行特异性加工的颞叶人声区(temporal voice areas, TVA) (即听觉皮层中对人类声音有选择性反应的区域)，沿双侧 STG/S 分布(Belin & Grosbras, 2010; Schirmer, 2018; Whitehead & Armony, 2018)。最近，有研究者采用功能磁共振人声加工区域定位分析技术(fMRI ‘voice localizer’)对 200 多名被试的 TVA 进行了快速且可靠的定位分析(Pernet et al., 2015)。研究结果表明大多数被试(94%)的双侧 STG/S 对人类声音比非人类声音的反应更加强烈。聚类分析进一步发现反应峰值的位置沿 STG/S 的前、中和后部分布。

尽管以上的研究表明人声加工与 TVA 的激活有着密切的关系，但是并不能说明两者存在因果关系。为了探究两者的因果关系，Bestelmeyer, Belin 和 Grosbras (2011)利用重复经颅磁刺激技术(repetitive transcranial magnetic stimulation, rTMS)分别对被试的右侧 TVA 和控制位置(右侧缘上回)进行刺激。结果发现当 rTMS 刺激控制位点时，被试在人声感知任务(即对人声和非人声进行分类)和响度识别任务(即判断声音的响度)中的表现水平在刺激前后没有出现显著变化；当 rTMS 刺激右侧 TVA 时，被试在人声感知任务中的表现水平较刺激前有所下降，而在响度识别任务中的表现水平在刺激前后仍然没有变化。这项研究首次表明 TVA 与人声加工之间存在着因果联系，进一步证明 TVA 是人声加工的特异性脑区。

2.2　电生理研究的证据

除了脑成像研究，电生理研究也证明了大脑对人类声音有特异性反应。Levy, Granot 和 Bentin (2001)采用 oddball 范式，让被试聆听 13 种乐器分别演奏的乐音和 4 名歌手分别唱出的乐音以及钢琴声，其中钢琴声作为靶刺激(概率 10%)，要求被试对靶刺激做出按键反应。ERP 结果发现相对乐器演奏的乐音，歌手唱出的乐音在声音刺激出现 320 ms 左右会诱发显著的正电位成分，这一成分被称为人声特异性脑电成分(VSR, voice-specific response)。Charest 等(2009)采用鸟声、人声和环境声音作为声音刺激，要求被试对声音类别进行辨别反应，并做出相应的按键反应。研究结果发现在刺激出现后 164~200 ms，人声在额−颞电极比其他种类的声音诱发出更大的正成分。这个脑电成分被称为“额−颞区正向电位” (fronto-temporal positivity to voice, FTPV)，它是人类声音激活 TVA 后产生的电生理成分，属于听觉的 P2 成分。由于 FTPV 与人脸识别的早期成分 N170 在时间进程上一致，所以研究者把 FTPV 称为听觉上的 N170。

这一成分反映了特异性的人声早期知觉加工。后来, Capilla (2013)等利用 MEG 技术进一步证明了 FTPVm (magnetic counterpart of the FTPV)的存在。他们让被试聆听一系列不同的声音刺激, 包括人声刺激(言语人声和非言语人声)和非人声刺激(动物发声、自然声音、人工合成声音), 并要求被试完成不同注意要求的任务(被动聆听任务、1-back 任务和对人声−非人声分类的任务)。结果表明在刺激呈现后 150 ms, 大脑就能够对人声和非人声进行区分, 并在 220 ms 左右 FTPVm 达到峰值。另外, 在三种不同的任务中, 人声都能诱发出明显的FTPVm成分, 且该成分来源于双侧STG/S 的中部, 大部分与 TVA 重叠。

2.3 来自婴儿的证据

4 个月左右的婴儿已经对人声有特异性反应了。研究者采用行为偏好范式探究了 3 个月内的婴儿的听觉偏好(Vouloumanos, Hauser, Werker, & Martin, 2010)。行为偏好范式是通过记录婴儿的吸吮次数来考察婴儿的行为偏好, 吸吮次数越多就表明偏好程度越高。最后的结果表明 3 个月内的婴儿比起合成声音更偏爱人类言语发声。随后, 研究者对刚出生 1~5 天的新生儿采用 oddball 范式进行了 ERP 实验, 结果发现相比于非人声刺激, 人声刺激诱发更大的"失匹配反应" (mismatch response, MMR), 而且恐惧、愤怒的人声比高兴的人声要诱发更大的 MMR (Cheng, Lee, Chen, Wang, & Decety, 2012)。这说明新生儿不仅能区分人声和非人声, 还能区分人声的情绪信息。Grossmann, Oberecker, Koch 和 Friederici 等(2010)采用 fNRIS 考察 4 个月和 7 个月的婴儿聆听人类声音(言语刺激和非言语刺激)和非人类声音时的大脑活动特点, 研究结果发现 7 个月婴儿的颞上皮层(superior temporal cortex, STC)对人声有显著的选择性反应; 而 4 个月大婴儿的颞上皮层没有表现出对人声的选择性反应。一些研究者认为 4 个月和 7 个月的婴儿对人声和非人声刺激的反应不同, 可能是因为 Grossmann 等的实验中选取的人声刺激存在问题(Lloyd-Fox, Blasi, Mercure, Elwell, & Johnson, 2012)。随后, Lloyd-Fox 等(2012)选取非言语人声刺激(如哭声、笑声、咳嗽声等)和熟悉的非人声刺激(如水流声、玩具的嘎嘎声)作为人声材料, 这就排除了言语和动物发声的干扰, 并控制了声音的熟悉度。fNRIS 结果表明 4~7 月婴儿的双侧前 STC 都对人声刺激比非人声刺激的反应更强烈, 且反应强度随着年龄增长而稳定增强(Lloyd-Fox et al., 2012)。鉴于 fNRIS 的空间分辨率不如 fMRI 高, Blasi 等采用 fMRI 技术对 3~7 月婴儿进行了研究, 最后观察到婴儿和成人相似, 右侧的前颞上回对非言语人声有选择性反应(Blasi et al., 2011)。

2.4 来自非人类动物的证据

不只是人类存在对人声有特异性反应的脑区(即 TVA), 其它物种也存在了类似于 TVA 的脑区。Petkov 等(2008)在 fMRI 实验中发现清醒猕猴的颞叶对猕猴发声比其他复杂声音有更强烈的反应, 并且右侧颞叶的前部可能参与了不同猕猴发声的识别。这个实验首次证明了猕猴有类似于人类 TVA 的大脑皮层。随后, 研究者对猕猴的声音选择性区域进行了单细胞记录, 结果发现了这些脑区确实存在对同物种发声有选择性反应的神经元(Perrodin, Kayser, Logothetis, & Petkov, 2011)。这些研究表明人声的特异性加工可能是进化的产物。有趣的是, 最近的研究发现狗的大脑中也存在类似于 TVA 的区域, 该区域对狗叫声比其他声音有更强烈的反应, 这表明人声的特异性加工脑区可能在 800 万年前就已经出现了(Andics & Miklósi, 2018; Andics, Gácsi, Faragó, Kis, & Miklósi, 2014)。

3 人声加工的神经机制

人声加工是以发声为基础的。人声是由声源(喉部的声带)和过滤器(喉部上方的声道)相互作用而产生的(Ghazanfar, & Rendall, 2008)。最常见的人声(浊音)是具有特定基频的声带的周期性振荡。个体在正常发音或唱歌时所达到的基频范围是相当宽泛的, 但是个体的平均基频是声带大小的函数, 例如男性比女性或小孩有更大的声带, 所以男声的基频值更低(Latinus, & Belin, 2011)。喉部上方的声道像一个滤波器, 使得在声源中的特定频率上产生共振(称为共振峰)。共振峰频率取决于发声器官的特定结构, 也取决于个体声道的大小(Latinus et al., 2011)。因此当发出同一个元音, 男性比女性或小孩具有更低的共振峰频率。发声器官结构的细微差异决定了说话者嗓音的独特性。值得注意的是, 除了正常发音方式(声道收缩的程度和类型)外, 喉部也能发出"假声"和"气泡音", 这就造成了人声的多样性。

人声的独特发声机制使得人声的声学特征不同于其他种类的声音。与其他种类的声音相比，人声的一个显著特征是特定共振峰的频率通常会快速地变化，反映了发声器官从一个位置移向另一个位置时声道形状的变化(Moore, 2008)。人声的另一个显著特征是更加谐和，即人声在时间频谱上比大多数声音类别更规律(除了乐器)。这种规律可以通过诸如谐波噪声比(harmonic-to-noise ratio)、基频微扰(jitter)和振幅微扰(shimmer)等指标来观测到，并且这些指标可以用于计量基频和振幅的短期微扰(Latinus et al., 2011)。此外，不同于其他种类声音，人声的声学特征还传递着重要的社会相关信息(Belin et al., 2004)。共振峰频率的变化传递着语言信息(一些语言除外，如普通话可以根据不同基频来识别)。基频携带着语言信息和情感韵律信息。音色就像视觉上的形状一样，是说话者身份识别的重要线索。

人声是由频率和强度随时间变化的声学模式组成的。当声音传入人耳，复杂的宽带声音能通过听觉过滤器分解为窄带信息，然后由希尔伯特变换(Hilbert transform)的形式进一步分解成变化速度较快的时间精细结构(emporal fine structure, TFS)成份和变化速度较慢的包络(envelope)成份(Moore, 2008)。TFS 在基频、言语的感知以及声源定向中起着重要作用，包络对声音的分类、音色的分析以及言语的可懂度至关重要，且这两种成分分别是形成“内容(what)”神经通路和“空间(where)”神经通路的声学基础(Apoux, Yoho, Youngdahl, & Healy, 2013; Zeng et al., 2004)。来自同一声音的包络信号和精细结构信号可以在知觉层面上捆绑成一个对应于该声源的特定听觉客体，这就使得听者能够在复杂的听觉环境中区分不同说话者的身份及其说话内容。

研究者认为人声和人脸一样主要传递着言语、情绪和身份信息，且三种信息加工的神经通路部分分离(Belin et al., 2004)。本文接下来将阐述人声言语、情绪和身份信息加工的神经机制。

3.1　人声言语信息加工的神经机制

人声中的言语信息加工是个体通过听觉通道接受声音流，感知其中的语音信息，并获得意义的过程。Belin 等(2004)结合过去 20 多年的研究提出大脑中存在专门加工人声言语的神经通路。目前，大量正常人和脑损伤病人的神经成像研究证明了大脑中存在专门加工言语信息的神经通路，并对该神经通路做出了深入探究(Hickok & Poeppel, 2016; Leonard, Cai, Babiak, Ren, & Chang, 2016; van der Burght, Goucha, Friederici, Kreitewolf, & Hartwigsen, 2019)。

语言是音和义的结合体。一些研究发现语音和语义的加工过程是相互分离的(Demonet et al., 1992; Okada, Matchin, & Hickok, 2018; Rong, Isenberg, Sun, & Hickok, 2018; Vaden Jr, Muftuler, & Hickok, 2010)。早在 20 世纪, Demonet 等(1992)为了将语音和语义的加工过程分离，用音节、音素等亚词汇的识别任务来考察语音加工，而用单词、句子等的识别任务来考察语义加工。他们要求被试分别进行音素识别任务和单词识别任务，最后的结果表明两种任务会激活不同的脑区。这就揭示了语音加工和语义加工涉及不同的脑区。此外，脑损伤病人的研究发现有些失语症患者的音节识别能力受损，但单词的语义理解能力完好；而有些患者的音节识别能力完好，但单词语义的理解能力受损(Dial & Martin, 2017)。这也进一步说明了语音和语义的加工可能是相互独立的。

在语音加工方面，研究者通常会通过操纵语音条件来调节语音加工的脑活动，如操纵词汇的相邻语音密度(phonological neighborhood density) (Okada & Hickok, 2006)。单词的相邻语音密度可由听起来和该单词相似的单词的数量(即改变该单词的一个音素后可获得新单词的数量)来测得，比如单词“rat”的相邻语音密度较高(cat、bat、hat、ram、rag、rap 等)，而单词“orange”的相邻语音密度较低。Okada 和 Hickok (2006)在 fMRI 实验中发现与低密度单词相比，被试在聆听高密度单词时，双侧 pSTS 会有更大的激活，这就表明 pSTS 在词汇的语音加工过程中起着重要作用。后来，Vaden 等(2010)通过操纵单词表中相同音素的数量(即音素重复程度)以考察语音加工的神经活动。实验者向 17 名被试呈现不同音素重复程度(低、中、高)的单词列表，最后观察到 STS 的中部(middle STS, mSTS)出现了明显的重复抑制效应，即该区域会随着语音重复程度的增高而反应降低。这一结果表明 mSTS 也参与了语音加工。这些研究表明 pSTS 和 mSTS 在语音加工中起着关键的作用。

关于语义加工的研究有很多。Rodd, Davis 和

Johnsrude (2005)在 fMRI 实验中要求被试聆听包含高模糊单词的句子和低模糊单词的句子。相对于低模糊单词，高模糊单词的加工还需要大脑对上下文相关词义进行激活和选择。研究结果发现高模糊单词比低模糊单词更能激活左侧颞下皮层的后部。这表明左侧颞下皮层的后部负责句子中的词义加工。此外，脑损伤病人的研究发现中风患者的单词理解障碍可能是左侧的后颞叶和颞中回损伤引起的(Bonilha et al., 2017)。这些研究表明词汇-语义加工可能涉及了左侧颞叶皮层的中后部。而相比于词义加工，当被试在对句子进行语义理解时，前颞叶(anterior temporal lobe, ATL)的反应更强烈(Brennan & Pylkkanen, 2017; den Ouden et al., 2019; Rice, Lambon Ralph, & Hoffman, 2015)。不过，ATL 在句子理解中起着何种作用迄今仍不清楚。一些研究支持 ATL 与句法结构的建立有关，如 ATL 的损伤会引起复杂句法结构的理解缺陷(Brennan & Pylkkaenen, 2012; den Ouden et al., 2019)。然而，原发性进行性失语症(primary progressive aphasia, PPA)的相关研究表明 ATL 与组合语义的加工有紧密的联系(Wilson et al., 2014)。综上所述，左侧颞叶的中后部是负责加工词汇语义的重要脑区，ATL 是句法结构和组合语义加工的神经网络中的重要脑区。

然而，语义和语音加工过程并不是完全独立的(Dial et.al., 2017; Dial, Mcmurray, & Martin, 2019)。研究发现威尔尼克失语症患者的语音感知和语义理解都存在缺陷，且语义理解障碍可能是由语音感知的缺陷所引起的，这就表明语义加工可能在一定程度上依赖于语音加工(Robson, Pilkington, Evans, DeLuca, & Keidel, 2017)。研究者推测语音和语义加工所激活的脑区很可能形成了一个神经回路，共同协作完成言语的加工过程(Hickok & Poeppel, 2007, 2016)。Hickok 等(2007)提出的双通路模型很好的解释了言语加工的脑机制。在双通路模型中，双侧 mSTS 和 pSTS 负责声音刺激的语音加工和表征。随后，该模型分出两条通路：一条是腹侧通路，它将基于声音的语音表征映射到意义表征上，即对言语信息进行意义理解。在该通路中，后外侧颞叶(posterior lateral temporal lobe)负责听觉刺激的词汇-语义访问，ATL 参与高级句法和复合语义加工。另一条是背侧通路，它将基于声音的语音表征映射到发声运动表征上，其功能是作为一个界面将 STS 编码的语音表征转换成运动区域(位于额下回)编码的发声运动表征。这个模型较全面得解释了人声言语加工的神经机制。

除此之外，言语不仅携带着音素、单词和句子等语言内容，还包含了说话者的身份信息。从进化的角度来说，人声的言语加工和身份识别都是从早期的人声加工能力中发展出来的，两者存在较密切的关系(Creel, & Bregman, 2011)。目前，一些研究证明了言语信息加工能影响人声身份的识别。例如，跨文化研究表明被试对母语说话者比非母语说话者的声音识别能力更强(Perrachione, Pierrehumbert, & Wong, 2009; Wester, 2012)。脑损伤病人的研究发现与正常个体相比，读写障碍者(因语音加工受损导致阅读能力障碍的患者)对母语说话者的声音识别能力有明显的损伤；但是正常个体和读写障碍者对非母语说话者的声音识别能力没有显著差异(Perrachione, Del Tufo, & Gabrieli, 2011)。这些研究证明人声识别依赖于语言能力。随后，Fleming, Giordano, Caldara 和 Belin (2014)的研究发现即使母语说话者无法理解言语中的语义内容，其对母语的身份识别能力还是要强于非母语说话者。这进一步说明人声的身份识别更依赖于言语的声学结构感知而不是语义理解能力。

3.2 人声情绪信息加工的神经机制

在日常交流中，人们能从变化的声学线索中提取情绪信息，进而推断出说话者的情绪状态。由于言语情绪韵律(speech prosody)既包含情绪信息又包含语义内容，两种人声信息可能会相互影响，且言语情绪韵律中特定的语言不利于跨文化的比较(Belin et al., 2011)。所以，研究者通常使用由情绪语调发出的无意义假词组成的非言语句子或非语言发声(如笑声、惊叫声)来考察人声情绪信息加工的特征(Belin et al., 2011)。Bestelmeyer, Rouger, DeBruine 和 Belin (2010)首次采用非言语人声的适应范式探索了人声情绪加工。适应是指在持续的刺激过程中，大脑会更偏向于对具有与刺激特征相反的刺激进行反应。研究者通常利用适应来隔离和扭曲某一神经群对特定属性的感知，从而揭示该神经群能对特定的刺激属性做出反应。在该研究中，被试对人声情绪(恐惧或愤怒)产生了适应效应，但是当人声的声学特性和情绪特性被夸大时，适应效应没有得到增强

(Bestelmeyer et al., 2010)。研究者认为人声情绪信息的适应效应不仅仅是声学特征的低层次适应引起的，也是由于人声情绪的神经表征的高层次适应。这就说明人声情绪加工可能涉及了一条独立的神经通路。后来, Schirmer 和 Gunter (2017)利用电生理技术发现了人声情绪的加工过程可能独立于其他人声信息加工。他们让被试聆听带有惊奇、中性情绪的人声刺激与非人声刺激，ERP 结果表明相比于非人声刺激，人声会诱发更大的 N1 和 P2 成分，而带有情绪的人声刺激还会诱发更大的晚期正成分。研究者认为大脑经过人声和非人声的区分之后会对人声中的情绪线索进行整合(Schirmer & Gunter, 2017)。

神经心理学研究表明右半球受损比左半球受损对个体识别人声情绪的能力的损害更大(Guranski & Podemski, 2015; Shamay-Tsoory, Tomer, Goldsher, Berger, & Aharon-Peretz, 2004)。右半球损伤的病人无法判断句子中表达的情绪意义，却能正常感知句子的内容；而左半球损伤的病人无法判断句子内容，却能够识别出句子中的情绪性表达(Patel et al., 2018; Ross & Monnot, 2011)。此外，大量神经成像研究发现人声的情绪韵律的识别会显著激活右侧额下皮层、右侧颞中回、右侧 STG 等右半球脑区(Friederici & Alter, 2004; Sammler, Grosbras, Anwander, Bestelmeyer, & Belin, 2015)。因此，许多研究者们认为专门负责加工人声情绪信息的神经网络位于右半球。

然而近年来，越来越多的脑成像研究表明情绪韵律加工可能不止涉及到右半球，还涉及广泛的双侧神经网络(Peg, Kotz, & Belin, 2017; Schirmer & Kotz, 2006; Ethofer et al., 2012; Zhang, Zhou, & Yuan, 2018)。Frühholz 和 Grandjean (2013)认为情绪声音能激活双侧额下皮层(inferior frontal cortex, IFC)，且左、右侧 IFC 的功能活动表现出相似的前后梯度变化。此外, IFC 不仅仅涉及情绪声音的注意加工和认知评价，还涉及对情绪声音的前注意加工和内隐加工。并且 IFC 的不同亚区具有不同的功能，头端腹侧额下回主要负责情绪声音的类别加工，而尾端背侧额下回主要加工情绪声音的时间特征信息(Frühholz et al., 2013)。Ethofer 等(2012)的弥散张量成像研究(diffusion tensor imaging)发现在情绪韵律识别中，双侧颞上回(STG)与其同侧的内侧膝状体(medial geniculate body, MGB)、双侧顶下叶(inferior parietal lobe, IPL)与其同侧的额下回(inferior frontal gyrus, IFG)具有较强的联结。双侧 STG 和其同侧 MGB 的联结反映了人声中情绪线索的早期输入，双侧 IPL 与其同侧 IFG 的联结反映了大脑在更高层次上对人声情绪信息和空间位置的加工(Ethofer et al., 2012; Zhang et al., 2018)。这说明双侧大脑皮层都参与了情绪声音的识别。

除此之外，研究发现人声的情绪加工可能还涉及皮层下结构，比如岛叶、杏仁核等(Bestelmeyer, Maurage, Rouger, Latinus, & Belin, 2014; Frühholz, Trost, & Kotz, 2016; Leitman, Edgar, Gamez, & Roberts, 2016)。Bestelmeyer 等(2014)使用声音的适应范式发现双侧 STS 和杏仁核对愤怒−恐惧连续刺激的物理声学特征差异更敏感，而前额区域和脑岛的前部对愤怒−恐惧连续刺激的情绪感知差异更敏感。这项研究表明除了额颞叶皮层，杏仁核和前脑岛也参与了人声的情绪加工，其中杏仁核负责分析情绪声音的声学特征，前脑岛负责人声情绪的认知表征。

Schirmer 和 Kotz (2006)提出的人声情绪加工的多阶段模型对情绪韵律的加工进行了很好的解释，并强调了人声情绪韵律的加工需要双侧神经网络协调进行。该模型把人声情绪加工分成感觉加工阶段、整合阶段和认知评价阶段。在声音刺激出现后的 100 ms 左右，初级和次级听觉皮层对输入刺激的声学信息(如振幅、时间、基频等)进行提取和分析。在刺激出现后的 200 ms 左右，前颞上沟或回(aSTS/G)和杏仁核对具有情绪意义的声学线索(如效价、唤醒、特定情绪特性等)进行整合。在刺激开始后的 400 ms 左右，右侧 IFG 和眶额皮质(orbitofrontal cortex, OFC)负责对情绪韵律进行更高级的认知评价，左侧额下皮层负责加工言语中的语义情绪信息。

3.3　人声身份信息加工的神经机制

由于声带、喉头等发声器官的结构特征存在个体差异，所以不同个体嗓音的声学参数也存在着细微的差异。有研究者认为大脑能对不同嗓音的独特声学特征进行感知分析，然后将输入的人声感知和储存在“人声识别单元(voice recognition units)”中的人声表征进行对比，最后识别出人声身份(Belin et al., 2004; Blank, Wieland, & von Kriegstein, 2014; Ellis, Jones, & Mosdell, 1997)。这

种传统观念强调人声身份的加工过程是按照从身份感知阶段到身份识别阶段的顺序进行的。通俗的来说，人声感知是指区分不同陌生说话者发出的声音，人声识别是指再认出熟悉的声音。然而，另外一些研究者反对人声身份的加工过程是按照这种严格的顺序进行的。脑损伤研究发现右颞叶肿瘤患者能正常识别熟悉的人声，却难以区分不熟悉的人声，这表明不熟悉声音的区分过程和熟悉声音的识别过程可能部分独立(Papagno, Mattavelli, Casarotti, Bello, & Gainotti, 2017)。此外，MEG 研究发现大脑在大约 200 ms 的时间点开始对不熟悉人声和熟悉人声同时进行反应，这就揭示了人声−身份的感知和识别是同时进行的(Schall, Kiebel, Maess, & von Kriegstein, 2015)。这些研究说明人声身份的感知过程和识别过程可能部分独立且并列存在。

脑损伤研究发现人声失认症(phonagnosic)患者能够理解人声中的情绪内容和言语含义，却不能通过声音识别出个体身份，这就表明人声身份的加工可能涉及了独立的神经通路(Roswandowitz et al., 2014)。一些研究者把人声身份加工的神经通路称为核心人声系统，该系统主要包括颞横回(heschl's gyrus, HG)、颞平面(planum temporale, PT)、颞上回/沟的前中后部以及部分颞中回/沟(Roswandowitz, Schelinski, &von Kriegstein, 2017; Schelinski, Borowiak, & von Kriegstein, 2016)。这些区域在人声−身份信息的加工过程中发挥着潜在的不同作用，并且在功能和结构上相互连接、相互作用(Roswandowitz et al., 2017)。

在核心人声系统中，颞横回(heschl's gyrus, HG)、颞平面(planum temporale, PT)和后颞上沟/回(pSTS/G)负责人声−身份的声学特征分析(Andermann, Patterson, Vogt, Winterstetter, & Rupp, 2017; von Kriegstein, Smith, Patterson, Kiebel, & Griffiths, 2010; Elmer, Hänggi, & Jäncke, 2016; Zäske, Hasan, & Belin, 2017)。例如, HG 对不同人声身份的音高变化更敏感(Andermann et al., 2017); pSTS/G 对人声的音色变化更为敏感(von Kriegstein et al., 2010); PT 和 pSTS 不仅对变化的人声身份更敏感，还对与人声身份线索有关的时频特性的变化更加敏感(Elmer et al., 2016)。人声加工的一个重要功能是区分不同陌生说话人的身份，而人声身份变化的感知和人声身份的声学特征分析密切相关。Zäske 等(2017)发现 pSTS/G 在不熟悉人声的区分过程中起了关键作用。他们先让被试学习一系列人声身份，然后要求被试判断呈现的人声是来自学习过的声音还是陌生的声音。最后的结果观察到 pSTS/G 对陌生声音的反应比熟悉声音的反应更强烈。人声中声学特征的提取(如频谱包络)和说话人身份变化的感知可能涉及了共同的脑区，即 pSTS/G。

研究者们还发现熟悉人声的身份识别是由 STS/G 前部到中部的脑区负责(Belin & Zatorre, 2003; Hasan, Valdessosa, Gross, & Belin, 2016; Luzzi et al., 2018; Schelinski et al., 2016)。Belin 等(2003)利用声音适应范式发现被试在适应同一说话者发出的不同音节后(即适应说话者的身份后), aSTS/G 反应强度下降，这就表明该区域对人声身份进行了识别。而被试在适应不同说话者发出的同一音节后(即适应言语后), aSTS/G 的反应没有减弱，这也进一步表明该区域可能只对人声身份进行加工，而对言语信息没有反应。另外，研究者向被试呈现一系列陌生人声的样本，这些人声样本被编辑成以线性的方式与原型声音偏离(Latinus, McAleer, Bestelmeyer, & Belin, 2013)。原型声音是在一个三维人声空间中将多个人声的基频(fundamental frequency)、共振峰分散(formant dispersion)和谐波噪声比(harmonics-to-noise ratio)等声学特征进行平均而构建起来的。结果发现识别偏离原型的声音比识别接近原型的声音更能激活 mSTS/G。因此，研究者认为 mSTS/G 可能参与了人声身份中独特的声学特征分析和身份识别之间的中间计算过程(Latinus et al., 2013)。换句话说，颞上回/沟的中部可能促进了人声身份加工过程中颞上回/沟的后部和前部之间功能连接(Roswandowitz, Kappes, Obrig, & von Kriegstein, 2017)。

为了解释人声身份信息的神经机制, Maguinness, Roswandowitz 和 von Kriegstein (2018)提出了整合模型。根据整合模型，人声身份信息在感知层面的加工，是对身份信息的声学特征进行提取与合并(即身份特征分析)，主要由后颞上沟/回(pSTS/G)、颞平面(PT)和前外侧颞横回(anterolateral HG)负责。在人声身份识别层面上，那些被提取的人声身份特征将在 STG/S 的中间区域中与已存储的人声原型进行比较，进而选择出偏离人声原型的特征。然后，aSTG/S 和 mSTS/G 会对偏离的特征

与“存储参照图式”进行比较，并计算出两者的差距，即参照图式比较(d)。存储的参照图式是每个人声身份所特有的，位于 aSTG/S。如果两种图式足够匹配，即“d”低于某个知觉阈值(Th)，人们就会产生一种熟悉感，即人声−身份识别(Fontaine, Love, & Latinus, 2017)。如果两者不匹配，人们会感觉到呈现的声音是陌生的，迭代循环会自动启动。这个迭代循环包括了语音身份特征分析和参照模式比较两个过程，参照图式是通过多次的迭代循环而建立起来的。

根据听−视觉整合模型，人声加工和人脸加工系统在多个加工阶段会产生交互作用(Maguinness et al., 2018)。这一观点得到了一些研究的支持。来自脑损伤病人的研究发现，相比正常人，发展性人脸失认症患者对于熟人声音的识别出现了障碍，但对于陌生人声音的识别却表现正常(Liu, Corrow, Pancaroglu, Duchaine, & Barton, 2015; von Kriegstein et al., 2008)。神经成像研究表明当被试对熟人的声音进行识别时，梭状回人脸识别区(FFA)和人声识别区(STG/S)会有较强的功能连接和结构连接(Schall, & von Kriegstein, 2014; von Kriegstein, Kleinschmidt, & Giraud, 2006; von Kriegstein, Kleinschmidt, Sterzer, & Giraud, 2005)。此外，人脸加工网络中的其他区域也参与人声识别加工。例如，Blank, Kiebel 和 von Kriegstein (2015)使用人声−人脸启动范式发现枕叶人脸识别区(occipital face area, OFA)对人声的物理特征和身份信息加工都敏感，前颞叶人脸识别区(anterior temporal lobe face area, aTL-FA)和 FFA 能对人声的身份信息进行表征。这些研究说明视觉人脸加工系统可能在人声身份识别中起着整合的作用。

4　研究展望

第一，人声加工的特异性问题还有待进一步探讨。首先，神经成像研究很难确定人脑中的特异性反应，并且研究者们对特异性神经反应所参照的标准至今仍没有达到共识(Pernet, Schyns, & Demonet, 2007; Leech, & Saygin, 2011)。以往研究通常把人声加工比其他类型声音加工有更强的神经激活看作是人声的特异性加工(Pernet et al., 2007; Leech, & Saygin, 2011)。不过，Pernet 等(2007)认为如果某一区域对一些声学刺激的神经反应都有增强，而对人声的反应更强一点，这种情况只能被认为是人声加工的优先性(voice-preferential)，而不是人声加工的特异性。尽管上颞区被传统地认为是人声特异性区域，但近年来，越来越多的研究发现上颞区不仅对韵律、音高等声学特征有强烈的反应，对乐器声、环境声等非人类声音也有强烈的反应(Armony, Aubé, Angulo-Perkins, Peretz, & Concha, 2015; Leaver et al., 2010; Leech, & Saygin, 2011)。因此，有研究者认为上颞区对人声的强烈激活反映得不是人声的特异性加工，而是对不同种类声音刺激进行的一般高级听觉加工过程，只不过对人声的敏感程度更高(Leech, & Saygin, 2011)。其次，Moerel, de Martino 和 Formisano (2012)的研究发现人声选择性区域对人声的低频特征有选择性反应，这表明听觉皮层中的人声选择性区域不能解释为人声加工的独立模块，同时也说明人声感知与一般声学机制之间可能存在着紧密联系。另外，Leaver 等(2010)发现听觉皮层对人声和非人声的频谱结构和时间调制特征具有选择性反应，并认为听觉皮层可能是根据声音中特定的频率和时间特征对声音进行分类。因此，人声加工的特异性可能是大脑对人声中特定的声学特征具有选择性的结果。

第二，未来还需对特殊群体的人声加工进行深入探究。目前，孤独症谱系障碍(autism spectrum disorders, ASD)者在 2 岁以前无法被确诊，这给 ASD 者的治疗和干预带来了限制。研究者们认为高风险 ASD 群体的前瞻性纵向研究有助于研究者找出 ASD 的治疗方法和提高 ASD 的干预效率(Jones, Gliga, Bedford, Charman, & Johnson, 2014; Sperdin & Schaer, 2016)。已有研究发现 ASD 儿童和 ASD 成人都不会对人声进行选择性反应，但关于高风险 ASD 婴儿的相关研究偏少(Bidet-Caulet et al., 2017; Charpentier et al., 2018; Fusaroli, Lambrechts, Bang, Bowler, & Gaigg, 2016)。Blasi 等(2015)发现无 ASD 家族遗传史的低风险婴儿能对人声进行选择性反应，而有 ASD 家族遗传史的高风险婴儿对人声和非人声的加工不存在显著差异。这一初步的研究表明异常的人声加工在未来可能成为 ASD 的确诊指标，但未来还需对此进行进一步验证。除此之外, ASD 儿童和 ASD 成人的听觉加工系统对声音刺激进行加工的早期阶段会出现功能和结构异常。比如，Edgar 等(2015)发现

和典型发育儿童相比, 6~14 岁 ASD 儿童的初级听觉区域会出现异常的发展趋势。Miron 等(2016)证明了异常听性脑干反应的 0~3 个月婴儿和 1.5~3.5 岁幼儿长大以后会被诊断为 ASD。未来研究还可以探究听觉系统的早期加工阶段异常如何影响自闭症患者的言语加工发展。

第三, 研究者们对自我面孔进行了大量的探究, 而对自我声音的关注相对较少。这是由于当人们说话时听到的自我声音既能通过空气传导又能通过骨传导, 但是自我声音的录音在实验中只能通过空气传导。近年来, 越来越多的研究发现自我录音也能诱发自我效应, 并且精神分裂症和自闭症患者对自我录音的加工出现异常, 所以越来越多的研究者开始关注自我声音加工的神经机制(Pinheiro, Farinha-Fernandes, Roberto, & Kotz, 2019; van Veluw & Chance, 2014)。但迄今为止, 研究者们只对自我声音加工的神经机制进行了初步的探索, 未来还需对以下几个方面进行深入探究。首先, 研究发现与自我声音相比较, 熟悉和陌生的他人声音都能诱发更大的 P3a (Graux et al., 2013; Graux, Gomot, Roux, Bonnet-Brilhault, & Bruneau, 2015)。这表明自我声音的识别过程可能不同于他人声音的识别过程 (Graux et al.,2013, 2015)。然而, 不同熟悉程度的声音(如名人的声音、朋友的声音、为了测试而学习过的声音、父母的声音或者兄弟姐妹的声音)可能会引起不同的神经活动(Graux et al., 2013)。为了排除自我声音可能属于某种类型的熟悉声音, 未来研究应对自我声音和不同熟悉程度的人声的加工过程进行深入的比较。其次, Graux 等(2013, 2015)发现在不注意条件下, 自我声音会比熟悉人声或陌生人声诱发更小的 P3a 成分, 而 Conde, Goncalves 和 Pinheiro (2015)发现在注意条件下, 自我声音比非自我声音诱发更大的 P3 振幅。这些研究表明自我声音和非自我声音加工受到注意资源的影响, 未来需深入探究注意资源如何调节自我声音和非自我声音的加工。另外, 研究者发现自我和非自我的音节声音能获得同等的注意资源, 而自我的单词声音比非自我的单词声音获得更多的注意资源(Conde, Goncalves, & Pinheiro, 2018)。这就表明自我人声加工的神经机制受到任务性质和刺激类型的影响, 未来研究可对这些影响进行考虑。最后, Pinheiro 等(2016)发现在言语加工过程中, 自我和他人声音对单词的语义情绪加工有不同的影响。随后, 他又发现与非自我声音相比, 神经分裂症患者对自我声音中的负性情绪内容更加敏感(Pinheiro et al., 2017)。未来需要在神经层面上进一步探究正常人和精神分裂症患者如何加工不同情绪类型的自我−他人声音。

参考文献

Agus, T. R., Paquette, S., Suied, C., Pressnitzer, D., & Belin, P. (2017). Voice selectivity in the temporal voice area despite matched low-level acoustic cues. *Scientific Reports, 7*(1), 11526.

Andermann, M., Patterson, R. D., Vogt, C., Winterstetter, L., & Rupp, A. (2017). Neuromagnetic correlates of voice pitch, vowel type, and speaker size in auditory cortex. *Neuroimage, 158*, 79–89.

Andics, A., Gácsi, M., Faragó, T., Kis, A., & Miklósi, Á. (2014). Voice-sensitive regions in the dog and human brain are revealed by comparative fMRI. *Current Biology, 24*(5), 574–578.

Andics, A., & Miklósi, Á. (2018). Neural processes of vocal social perception: Dog-human comparative fMRI studies. *Neuroscience & Biobehavioral Reviews, 85*, 54–64.

Apoux, F., Yoho, S. E., Youngdahl, C. L., & Healy, E. W. (2013). Role and relative contribution of temporal envelope and fine structure cues in sentence recognition by normal-hearing listeners. *The Journal of the Acoustical Society of America*, *134*(3), 2205–2212.

Armony, J. L., Aubé, W., Angulo-Perkins, A., Peretz, I., & Concha, L. (2015). The specificity of neural responses to music and their relation to voice processing: An fMRI-adaptation study. *Neuroscience Letters, 593*, 35–39.

Belin, P. (2017). Similarities in face and voice cerebral processing. *Visual Cognition, 25*(4-6), 658–665.

Belin, P., Bestelmeyer, P. E. G., Latinus, M., & Watson, R. (2011). Understanding voice perception. *British Journal of Psychology, 102*(4), 711–725.

Belin, P., Bodin, C., & Aglieri, V. (2018). A “voice patch” system in the primate brain for processing vocal information? *Hearing Research, 366*, 65–74.

Belin, P., Fecteau, S., & Bédard, C. (2004). Thinking the voice: Neural correlates of voice perception. *Trends in Cognitive Sciences, 8*(3), 129–135.

Belin, P., & Grosbras, M. H. (2010). Before speech: Cerebral voice processing in infants. *Neuron, 65*(6), 733–735.

Belin, P., & Zatorre, R. J. (2003). Adaptation to speaker's voice in right anterior temporal lobe. *Neuroreport, 14*(16), 2105–2109.

Belin, P., Zatorre, R. J., Lafaille, P., Ahad, P., & Pike, B. (2000). Voice-selective areas in human auditory cortex. *Nature, 403*(6767), 309–312.

Besson, G., Barragan-Jason, G., Thorpe, J., Fabre-Thorpe, M., Puma, S., Ceccaldi, M., & Barbeau, E. J. (2017). From face processing to face recognition: Comparing three different processing levels. *Cognition, 158*, 33–43.

Bestelmeyer, P. E. G., Belin, P., & Grosbras, M. H. (2011). Right temporal TMS impairs voice detection. *Current Biology, 21*(20), R838–R839.

Bestelmeyer, P. E. G., Maurage, P., Rouger, J., Latinus, M., & Belin, P. (2014). Adaptation to vocal expressions reveals multistep perception of auditory emotion. *Journal of Neuroscience, 34*(24), 8098–8105.

Bestelmeyer, P. E. G., Rouger, J., DeBruine, L. M., & Belin, P. (2010). Auditory adaptation in vocal affect perception. *Cognition, 117*(2), 217–223.

Bidet-Caulet, A., Latinus, M., Roux, S., Malvy, J., Bonnet-Brilhault, F., & Bruneau, N. (2017). Atypical sound discrimination in children with ASD as indicated by cortical ERPs. *Journal of Neurodevelopmental Disorders, 9*(1), 13.

Blank, H., Kiebel, S. J., & von Kriegstein, K. (2015). How the human brain exchanges information across sensory modalities to recognize other people. *Human Brain Mapping, 36*(1), 324–339.

Blank, H., Wieland, N., & von Kriegstein, K. (2014). Person recognition and the brain: Merging evidence from patients and healthy individuals. *Neuroscience & Biobehavioral Review, 47*, 717–734.

Blasi, A., Mercure, E., Lloyd-Fox, S., Thomson, A., Brammer, M., Sauter, D., ... Deoni, S. (2011). Early specialization for voice and emotion processing in the infant brain. *Current Biology, 21*(14), 1220–1224.

Blasi, A., Lloyd-Fox, S., Sethna, V., Brammer, M. J., Mercure, E., Murray, L., ... Johnson, M. H. (2015). Atypical processing of voice sounds in infants at risk for autism spectrum disorder. *Cortex, 71*, 122–133.

Bonilha, L., Hillis, A. E., Hickok, G., den Ouden, D. B., Rorden, C., & Fridriksson, J. (2017). Temporal lobe networks supporting the comprehension of spoken words. *Brain, 140*(9), 2370–2380.

Brennan, J., & Pylkkaenen, L. (2012). The time-course and spatial distribution of brain activity associated with sentence processing. *Neuroimage, 60*(2), 1139–1148.

Brennan, J. R., & Pylkkanen, L. (2017). MEG evidence for incremental sentence composition in the anterior temporal lobe. *Cognitive Science, 41*(S6), 1515–1531.

Bruce, V., & Young, A. (1986). Understanding face recognition. *British Journal of Psychology, 77*(3), 305–327.

Caharel, S., Montalan, B., Fromager, E., Bernard, C., Lalonde, R., & Mohamed, R. (2011). Other-race and inversion effects during the structural encoding stage of face processing in a race categorization task: An event-related brain potential study. *International Journal of Psychophysiology, 79*(2), 266–271.

Capilla, A., Belin, P., & Gross, J. (2013). The early spatio-temporal correlates and task independence of cerebral voice processing studied with MEG. *Cerebral Cortex, 23*(6), 1388–1395.

Charest, I., Pernet, C. R., Rousselet, G. A., Quiñones, I., Latinus, M., Fillion-Bilodeau, S., ... Belin, P. (2009). Electrophysiological evidence for an early processing of human voices. *BMC Neuroscience, 10*(1), 127.

Charpentier, J., Kovarski, K., Houy-Durand, E., Malvy, J., Saby, A., Bonnet-Brilhault, F., ... Gomot, M. (2018). Emotional prosodic change detection in autism spectrum disorder: An electrophysiological investigation in children and adults. *Journal of Neurodevelopmental Disorders, 10*(1), 28.

Cheng, Y., Lee, S.-Y., Chen, H.-Y., Wang, P.-Y., & Decety, J. (2012). Voice and emotion processing in the human neonatal brain. *Journal of Cognitive Neuroscience, 24*(6), 1411–1419.

Conde, T., Gonçalves, Ó. F., & Pinheiro, A. P. (2015). Paying attention to my voice or yours: An ERP study with words. *Biological Psychology, 111*, 40–52.

Conde, T., Goncalves, O. F., & Pinheiro, A. P. (2018). Stimulus complexity matters when you hear your own voice: Attention effects on self-generated voice processing. *International Journal of Psychophysiology, 133*, 66–78.

Creel, S. C., & Bregman, M. R. (2011). How talker identity relates to language processing. *Language and Linguistics Compass, 5*(5), 190–204.

Demonet, J. F., Chollet, F., Ramsay, S., Cardebat, D., Nespoulous, J. L., Wise, R., ... Frackowiak, R. (1992). The anatomy of phonological and semantic processing in normal subjects. *Brain, 115*(6), 1753–1768.

den Ouden, D.-B., Malyutina, S., Basilakos, A., Bonilha, L., Gleichgerrcht, E., Yourganov, G., ... Fridriksson, J. (2019). Cortical and structural-connectivity damage correlated with impaired syntactic processing in aphasia. *Human Brain Mapping, 40*(7), 2153–2173.

Dial, H., & Martin, R. (2017). Evaluating the relationship between sublexical and lexical processing in speech perception: Evidence from aphasia. *Neuropsychologia, 96*, 192–212.

Dial, H. R., Mcmurray, B., & Martin, R. C. (2019). Lexical processing depends on sublexical processing: Evidence from the visual world paradigm and aphasia. *Attention, Perception, & Psychophysics, 81*, 1047–1064.

Edgar, J. C., Fisk IV, C. L. F., Berman, J. I., Chudnovskaya, D., Liu, S., Pandey, J., ... Roberts, T. P. L. (2015). Auditory encoding abnormalities in children with autism spectrum disorder suggest delayed development of auditory cortex. *Molecular Autism, 6*(1), 69.

Ellis, H. D., Jones, D. M., & Mosdell, N. (1997). Intra- and inter-modal repetition priming of familiar faces and voices. *British Journal of Psychology, 88*(1), 143–156.

Elmer, S., Hänggi, J., & Jäncke, L. (2016). Interhemispheric transcallosal connectivity between the left and right planum temporale predicts musicianship, performance in temporal speech processing, and functional specialization. *Brain Structure & Function, 221*(1), 331–344.

Fecteau, S., Armony, J. L., Joanette, Y., & Belin, P. (2004). Is voice processing species-specific in human auditory cortex? An fMRI study. *Neuroimage, 23*(3), 840–848.

Fischer, J. (2017). Primate vocal production and the riddle of language evolution. *Psychonomic Bulletin & Review, 24*(1), 72-78.

Fleming, D., Giordano, B. L., Caldara, R., & Belin, P. (2014). A language-familiarity effect for speaker discrimination without comprehension. *Proceedings of the National Academy of Sciences, 111*(38), 13795–13798.

Fontaine, M., Love, S. A., & Latinus, M. (2017). Familiarity and voice representation: From acoustic-based representation to voice averages. *Frontiers in Psychology, 8*, 1180.

Friederici, A. D., & Alter, K. (2004). Lateralization of auditory language functions: a dynamic dual pathway model. *Brain and Language, 89*(2), 267–276.

Frühholz, S., & Grandjean, D. (2013). Processing of emotional vocalizations in bilateral inferior frontal cortex. *Neuroscience & Biobehavioral Reviews, 37*(10), 2847–2855.

Frühholz, S., Trost, W., & Kotz, S. A. (2016). The sound of emotions—Towards a unifying neural network perspective of affective sound processing. *Neuroscience & Biobehavioral Reviews, 68*, 96–110.

Fusaroli, R., Lambrechts, A., Bang, D., Bowler, D. M., & Gaigg, S. B. (2016). Is voice a marker for autism spectrum disorder? A systematic review and meta-analysis. *Autism Research, 10*(3), 384-407.

Ghazanfar, A. A., & Rendall, D. (2008). Evolution of human vocal production. *Current Biology, 18*(11), R457–R460.

Graux, J., Gomot, M., Roux, S., ... Camus, V. (2013). My voice or yours? An electrophysiological study. *Brain Topography, 26*(1), 72–82.

Graux, J., Gomot, M., Roux, S., Bonnet-Brilhault, F., & Bruneau, N. (2015). Is my voice just a familiar voice? An electrophysiological study. *Social Cognitive & Affective Neuroscience, 10*(1), 101–105.

Grossmann, T., Oberecker, R., Koch, S. P., & Friederici, A. D. (2010). The developmental origins of voice processing in the human brain. *Neuron, 65*(6), 852–858.

Guranski, K., & Podemski, R. (2015). Emotional prosody expression in acoustic analysis in patients with right hemisphere ischemic stroke. *Neurologia i Neurochirurgia Polska, 49*(2), 113–120.

Hasan, B. A. S., Valdessosa, M., Gross, J., & Belin, P. (2016). “Hearing faces and seeing voices”: Amodal coding of person identity in the human brain. *Scientific Reports, 108*(37494), 44.

Hickok, G., & Poeppel, D. (2007). The cortical organization of speech processing. *Nature Reviews Neuroscience, 8*(5), 393–402.

Hickok, G., & Poeppel, D. (2016). Neural basis of speech perception. *Handbook of Clinical Neurology, 129*, 149–160.

Jiang, X., Chevillet, M. A., Rauschecker, J. P., & Riesenhuber, M. (2018). Training humans to categorize monkey calls: Auditory feature- and category-selective neural tuning changes. *Neuron, 98*(2), 405–416.

Jones, E. J. H., Gliga, T., Bedford, R., Charman, T., & Johnson, M. H. (2014). Developmental pathways to autism: A review of prospective studies of infants at risk. *Neuroscience and Biobehavioral Reviews, 39*, 1–33.

Kanwisher, N., McDermott, J., & Chun, M. M. (1997). The fusiform face area: A module in human extrastriate cortex specialized for face perception. *The Journal of neuroscience, 17*(11), 4302–4311.

Latinus, M., & Belin, P. (2011). Human voice perception. *Current Biology, 21*(4), R143–R145.

Latinus, M., McAleer, P., Bestelmeyer, P. E. G., & Belin, P. (2013). Norm-based coding of voice identity in human auditory cortex. *Current Biology, 23*(12), 1075–1080.

Leaver, A. M., & Rauschecker, J. P. (2010). Cortical representation of natural complex sounds: Effects of acoustic features and auditory object category. *Journal of Neuroscience, 30*(22), 7604–7612.

Leech, R., & Saygin, A. P. (2011). Distributed processing and cortical specialization for speech and environmental sounds in human temporal cortex. *Brain and Language, 116*(2), 83–90.

Leitman, D. I., Edgar, C., Berman, J., Gamez, K., Fruhholz, S., & Roberts, T. P. (2016). Amygdala and insula contributions to dorsal-ventral pathway integration in the prosodic neural network. *arXiv preprint arXiv, 1611*, 01643.

Leonard, M. K., Cai, R., Babiak, M. C., Ren, A., & Chang, E. F. (2016). The peri-Sylvian cortical network underlying single word repetition revealed by electrocortical stimulation and direct neural recordings. *Brain & Language, 193*,

58–72.

Levy, D. A., Granot, R., & Bentin, S., (2001). Processing specificity for human voice stimuli: Electrophysiological evidence. *Neuroreport, 12*(12), 2653–2657.

Lewis, J. W., Talkington, W. J., Walker, N. A., Spirou, G. A., Jajosky, A., Frum, C., & Brefczynski-Lewis, J. A. (2009). Human cortical organization for processing vocalizations indicates representation of harmonic structure as a signal attribute. *Journal of Neuroscience, 29*(7), 2283–2296.

Liu, R. R., Corrow, S. L., Pancaroglu, R., Duchaine, B., & Barton, J. J. (2015). The processing of voice identity in developmental prosopagnosia. *Cortex, 71*, 390–397.

Lloyd-Fox, S., Blasi, A., Mercure, E., Elwell, C. E., & Johnson, M. H. (2012). The emergence of cerebral specialization for the human voice over the first months of life. *Social Neuroscience, 7*(3), 317–330.

Luzzi, S., Coccia, M., Polonara, G., Reverberi, C., Ceravolo, G., Silvestrini, M., ... Gainotti, G. (2018). Selective associative phonagnosia after right anterior temporal stroke. *Neuropsychologia, 116*, 154–161.

Maguinness, C., Roswandowitz, C., & von Kriegstein, K. (2018). Understanding the mechanisms of familiar voice-identity recognition in the human brain. *Neuropsychologia, 116*, 179–193.

Miron, O., Ari-Even, R. D., Gabis, L. V., Henkin, Y., Shefer, S., Dinstein, I., & Geva, R. (2016). Prolonged auditory brainstem responses in infants with autism. *Autism Research, 9*(6), 689–695.

Moerel, M., de Martino, F., & Formisano, E. (2012). Processing of natural sounds in human auditory cortex: Tonotopy, spectral tuning, and relation to voice sensitivity. *Journal of Neuroscience, 32*(41), 14205–14216.

Moore, B. C. J. (2008). The role of temporal fine structure processing in pitch perception, masking, and speech perception for normal-hearing and hearing-impaired people. *Journal of the Association for Research in Otolaryngology, 9*(4), 399–406.

Navajas, J., Nitka, A. W., & Quiroga, R. Q. (2017). Dissociation between the neural correlates of conscious face perception and visual attention. *Psychophysiology, 54*(8), 1138–1150.

Okada, K., & Hickok, G. (2006). Identification of lexical-phonological networks in the superior temporal sulcus using functional magnetic resonance imaging. *Neuroreport, 17*(12), 1293–1296.

Okada, K., Matchin, W., & Hickok, G. (2018). Phonological feature repetition suppression in the left inferior frontal gyrus. *Journal of Cognitive Neuroscience, 30 (*10), 1549–1557.

Papagno, C., Mattavelli, G., Casarotti, A., Bello, L., & Gainotti, G. (2017). Defective recognition and naming of famous people from voice in patients with unilateral temporal lobe tumours. *Neuropsychologia, 116*, 194–204.

Patel, S., Oishi, K., Wright, A., Sutherland-Foggio, H., Saxena, S., Sheppard, S. M., & Hillis, A. E. (2018). Right hemisphere regions critical for expression of emotion through prosody. *Frontiers in Neurology, 9*, 224.

Peg, B., Kotz, S. A., & Belin, P. (2017). Effects of emotional valence and arousal on the voice perception network. *Social Cognitive & Affective Neuroscience, 12*(8), 1351–1358.

Pernet, C. R., Mcaleer, P., Latinus, M., Gorgolewski, K. J., Charest, I., Bestelmeyer, P. E. G., ... Valdes-Sosa, M. (2015). The human voice areas: Spatial organization and inter-individual variability in temporal and extra-temporal cortices. *Neuroimage, 119*, 164–174.

Pernet, C., Schyns, P. G., & Demonet, J. F. (2007). Specific, selective or preferential: Comments on category specificity in neuroimaging. *Neuroimage, 35*(3), 991–997.

Perrachione, T. K., Del Tufo, S. N., & Gabrieli, J. D. (2011). Human voice recognition depends on language ability. *Science, 333*(6042), 595–595.

Perrachione, T. K., Pierrehumbert, J. B., & Wong, P. C. M. (2009). Differential neural contributions to native- and foreign-language talker identification. *Journal of Experimental Psychology: Human Perception and Performance, 35*(6), 1950–1960.

Perrodin, C., Kayser, C., Abel, T. J., Logothetis, N. K., & Petkov, C. I. (2015). Who is that? Brain networks and mechanisms for identifying individuals. *Trends in Cognitive Sciences, 19*(12), 783–796.

Perrodin, C., Kayser, C., Logothetis, N. K., & Petkov, C. I. (2011). Voice cells in the primate temporal lobe. *Current Biology, 21*(16), 1408–1415.

Petkov, C. I., Kayser, C., Steudel, T., Whittingstall, K., Augath, M., & Logothetis, N. K. (2008). A voice region in the monkey brain. *Nature Neuroscience, 11*(3), 367–374.

Pinheiro, A. P., Farinha-Fernandes, A., Roberto, M. S., & Kotz, S. A. (2019). Self-voice perception and its relationship with hallucination predisposition. *Cognitive Neuropsychiatry, 24*(4), 1–19.

Pinheiro, A. P., Rezaii, N., Rauber, A., Nestor, P. G., Spencer, K. M., & Niznikiewicz, M. (2017). Emotional self-other voice processing in schizophrenia and its relationship with hallucinations: ERP evidence. *Psychophysiology, 54*(9), 1252–1265.

Rice, G. E., Lambon Ralph, M. A., & Hoffman, P. (2015). The roles of left versus right anterior temporal lobes in conceptual knowledge: An ALE meta-analysis of 97 functional neuroimaging studies. *Cerebral Cortex, 25*(11),

4374–4391.

Robson, H., Pilkington, E., Evans, L., DeLuca, V., & Keidel, J. L. (2017). Phonological and semantic processing during comprehension in Wernicke's aphasia: An N400 and Phonological Mapping Negativity Study. *Neuropsychologia, 100*, 144–154.

Rodd, J. M., Davis, M. H., & Johnsrude, I. S. (2005). The neural mechanisms of speech comprehension: fMRI studies of semantic ambiguity. *Cerebral Cortex, 15*(8), 1261–1269.

Rong, F., Isenberg, A. L., Sun, E., & Hickok, G. (2018). The neuroanatomy of speech sequencing at the syllable level. *PloS One, 13*(10), e0196381.

Ross, E. D., & Monnot, M. (2011). Affective prosody: What do comprehension errors tell us about hemispheric lateralization of emotions, sex and aging effects, and the role of cognitive appraisal. *Neuropsychologia, 49*(5), 866–877.

Roswandowitz, C., Kappes, C., Obrig, H., & von Kriegstein, K. (2017). Obligatory and facultative brain regions for voice-identity recognition. *Brain, 141*(1), 234–247.

Roswandowitz, C., Mathias, S. R., Hintz, F., Kreitewolf, J., Schelinski, S., & von Kriegstein, K. (2014). Two cases of selective developmental voice-recognition impairments. *Current Biology, 24*(19), 2348–2353.

Roswandowitz, C., Schelinski, S., & von Kriegstein, K. (2017). Developmental phonagnosia: Linking neural mechanisms with the behavioural phenotype. *Neuroimage, 155*, 97–112.

Sammler, D., Grosbras, M. H., Anwander, A., Bestelmeyer, P. G., & Belin, P. (2015). Dorsal and ventral pathways for prosody. *Current Biology, 25*(23), 3079–3085.

Schall, S., Kiebel, S. J., Maess, B., & von Kriegstein, K. (2015). Voice identity recognition: Functional division of the right STS and its behavioral relevance. *Journal of Cognitive Neuroscience, 27*(2), 280–291.

Schall, S., & von Kriegstein, K. (2014). Functional connectivity between face-movement and speech-intelligibility areas during auditory-only speech perception. *PloS One, 9*(1), e86325.

Schelinski, S., Borowiak, K., & von Kriegstein, K. (2016). Temporal voice areas exist in autism spectrum disorder but are dysfunctional for voice identity recognition. *Social Cognitive and Affective Neuroscience, 11*(11), 1812–1822.

Schirmer, A. (2018). Is the voice an auditory face? An ALE meta-analysis comparing vocal and facial emotion processing. *Social Cognitive and Affective Neuroscience, 13*(1), 1–13.

Schirmer, A., & Gunter, T. C. (2017). Temporal signatures of processing voiceness and emotion in sound. *Social Cognitive and Affective Neuroscience, 12*(6), 902–909.

Schirmer, A., & Kotz, S. A. (2006). Beyond the right hemisphere: Brain mechanisms mediating vocal emotional processing. *Trends in Cognitive Sciences, 10*(1), 24–30.

Schroeder, J., Kardas, M., & Epley, N. (2017). The humanizing voice: Speech reveals, and text conceals, a more thoughtful mind in the midst of disagreement. *Psychological Science, 28*(12), 1745–1762.

Shamay-Tsoory, S. G., Tomer, R., Goldsher, D., Berger, B. D., & Aharon-Peretz, J. (2004). Impairment in cognitive and affective empathy in patients with brain lesions: anatomical and cognitive correlates. *Journal of Clinical and Experimental Neuropsychology, 26*(8), 1113–1127.

Sperdin, H. F., & Schaer, M. (2016). Aberrant development of speech processing in young children with autism: New insights from neuroimaging biomarkers. *Frontiers in Neuroscience, 10*, 393.

Vaden Jr, K. I., Muftuler, L. T., & Hickok, G. (2010). Phonological repetition-suppression in bilateral superior temporal sulci. *Neuroimage, 49*(1), 1018–1023.

van der Burght, C. L., Goucha, T., Friederici, A. D., Kreitewolf, J., & Hartwigsen, G. (2019). Intonation guides sentence processing in the left inferior frontal gyrus. *Cortex, 117*, 122–134.

van Veluw, S. J., & Chance, S. A. (2014). Differentiating between self and others: An ALE meta-analysis of fMRI studies of self-recognition and theory of mind. *Brain Imaging and Behavior, 8*(1), 24–38.

von Kriegstein, K., Dogan, Ö., Grüter, M., Giraud, A. L., Kell, C. A., Grüter, T., ... Kiebel, S. J. (2008). Simulation of talking faces in the human brain improves auditory speech recognition. *Proceedings of the National Academy of Sciences, 105*(18), 6747–6752.

von Kriegstein, K., Kleinschmidt, A., & Giraud, A. L. (2006). Voice recognition and cross-modal responses to familiar speakers' voices in prosopagnosia. *Cerebral Cortex, 16*(9), 1314–1322.

von Kriegstein, K., Kleinschmidt, A., Sterzer, P., & Giraud, A. L. (2005). Interaction of face and voice areas during speaker recognition. *Journal of Cognitive Neuroscience, 17*(3), 367–376.

von Kriegstein, K., Smith, D. R. R., Patterson, R. D., Kiebel, S. J., & Griffiths, T. D. (2010). How the human brain recognizes speech in the context of changing speakers. *Journal of Neuroscience, 30*(2), 629–638.

Vouloumanos, A., Hauser, M. D., Werker, J. F., & Martin, A. (2010). The tuning of human neonates' preference for speech. *Child Development, 81*(2), 517–527.

Wester, M. (2012). Talker discrimination across languages. *Speech Communication, 54*(6), 781–790.

Whitehead, J. C., & Armony, J. L. (2018). Singing in the brain: Neural representation of music and voice as revealed

by fMRI. *Human Brain Mapping, 39*(12), 4913–4924.
Wilson, S. M., Demarco, A. T., Henry, M. L., Gesierich, B., Babiak, M., Mandelli M. L., ... Gorno-Tempini, M. L. (2014). What role does the anterior temporal lobe play in sentence-level processing? Neural correlates of syntactic processing in semantic variant primary progressive aphasia. *Journal of Cognitive Neuroscience, 26*(5), 970–985.
Zäske, R., Hasan, B. A. S., & Belin, P. (2017). It doesn't matter what you say: fMRI correlates of voice learning and recognition independent of speech content. *Cortex, 94*, 100–112.
Zeng, F. G., Nie, K., Liu, S., Stickney, G., Del Rio, E., Kong, Y. Y., & Chen, H. (2004). On the dichotomy in auditory perception between temporal envelope and fine structure cues (l). *The Journal of the Acoustical Society of America, 116*(3), 1351.
Zhang, D., Zhou, Y., & Yuan, J. (2018). Speech prosodies of different emotional categories activate different brain regions in adult cortex: an fNIRS study. *Scientific Reports, 8*(1), 218.

Neural mechanisms for voice processing

WU Ke[1,2]; CHEN Jie[1,2]; LI Wenjie[1,2]; CHEN Jiejia[1,2]; LIU Lei[3]; LIU Cuihong[1,2]

([1] *School of Education Science, Hunan Normal University;*
[2] *Cognition and Human Behavior Key Laboratory of Hunan Province, Hunan Normal University, Changsha* 410081, *China*)
([3] *School of Psychological and Cognitive Sciences, Peking University, Beijing* 100080, *China*)

Abstract: The human voice is the most familiar and important sound in the human auditory environment, conveying large amounts of socially relevant information. Similar to face processing, there is also a functional specialization in brain for voice processing. Neuroimaging and electrophysiology studies have demonstrated that the temporal voice areas (TVAs) showed specific response to human voices. In addition, researchers have also observed the homologues of TVAs in non-human brain. Human voices can convey speech, affective and identity information, which are extracted and further processed in three interacting but partially dissociated neural pathways. To explicate these three functional pathways, researchers have proposed three corresponding models including the dual-stream model of speech processing, multi-stage model of vocal emotional processing and integrative model of voice-identity processing. In the future, researchers should further investigate whether voice-selective activity can be explained by the selective processing of specific acoustic features of voice and focus on neural mechanisms of voice processing in special populations (e.g. schizophrenia and autism).

Key words: voice processing; specialization; the temporal voice areas (TVA); speech processing; emotional prosody; voice-identity recognition